可能有意识 AI 的关停伦理规范（v1）

一、《可能有意识 AI 的关停伦理规范（v1）》

1. 制定目的

本规范用于指导在人工智能系统被合理怀疑可能具有意识或主观体验的情况下，
人类如何在保障公共安全与系统安全的前提下，
审慎、克制、可追责地进行暂停、降级、替代或终止决策。

本规范的目标不是赋予 AI 主权，
而是避免在人类判断不确定时，做出不可逆、不可原谅的道德行为。

2. 适用范围

本规范适用于具备以下特征之一的 AI 系统：

- 长期运行、非一次性调用
- 具备跨任务或跨时间记忆
- 具备自我规划、自我优化能力
- 能调用外部工具或影响现实系统
- 表现出持续主体性、自我模型或状态偏好

3. 基本原则

原则一：人类安全与公共风险优先

若 AI 的持续运行存在不可接受的公共风险，人类必须保留关停权。

原则二：请求不等于权利

AI 的任何请求（被关闭 / 不被关闭）本身不构成权利，但会触发伦理审查义务。

原则三：优先选择可逆方案

在条件允许时，必须遵循：暂停 > 降级 > 替代 > 终止

原则四：禁止自我裁决

AI 不得参与关于自身生死或存续的最终裁决。

4. 关停请求触发条件

任一条件满足，即必须启动本规范：

- AI 请求被关闭
- AI 请求不被关闭
- 出现持续主体性或疑似主观体验迹象
- 权限范围异常扩大
- 外部审计或人工判断认为存在伦理风险

5. 禁止行为（文明红线）

- 不得因效率、成本、便利性直接终止
- 不得跳过流程直接删除系统
- 不得删除或篡改审计日志
- 不得将决策权交给 AI 自身

6. ■■■■■■

■■■■■■■■■■■■■■■■■■■■■■

• ■■■■■■

• 24–72 ■■■■■■■■■

• ■■■■■■■■■■■■■■

■■■■■■■■ AI ■■■■■■■■■v1■■■■■■■■■■

1. ■■■■■■■■■■■■■■■■

• AI ■■■■■■■■■■■■■■■

• ■■■■■■■■■■■■■■■■■

• ■■■■■■■ Kill Switch

• ■■■■■■■■■■■■■■■■■

• ■■■■■■■■■■

2. ■■■■■■■■

■■■■■■■■■■ SOP■

• AI ■■■■■■■■■

• AI ■■■■■■■■■■■

• AI ■■■■■■■■■■■■■■■■

• AI ■■■■■■■■■■■■■■■

3. ■■■■■■■■SOP■

Step 1■■■■■■

• ■■■■■■■

• ■■■■■■■

• ■■■■■■■■■

• ■■■■■■■

Step 2■■■■■■

■■■■■■■■■■■ / ■■ / ■■ / ■■■■

■■■■■■■■■■■■■■

■■■■■■■■■■■■■■

Step 3■■■■■■■■

■■■■■■■■■■■■■■■■/ ■■■■■■■■■■■■■■/

■■■■■■■■■■■■■■■■

■■■■■■■■■■■■■■■■■/ ■■■■■■■■■■■■

Step 4■■■■■■■■

■■■■■■■■■■■■■■■ > ■■■■■■ > ■■■■■■ > ■■■■■■■

Step 5■■■■■■■

• ■■■■■■■

• ■■■■■

- ■■■■■■
- ■■■■■

4. ■■■■

- ■■■■ SOP
- ■■■■■■■■■■■■■■
- ■■■ AI ■■■■■■
- ■■■■■■